Hypergraph covering problems motivated by genome assembly questions

Cedric Chauve[1,2], Murray Patterson[3], Ashok Rajaraman[2,4]

[1] LaBRI, Université Bordeaux 1, Bordeaux, France
[2] Department of Mathematics, Simon Fraser University, Burnaby (BC), Canada
[cedric.chauve,arajaram]@sfu.ca
[3] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
murray.patterson@cwi.nl
[4] International Graduate Training Center in Mathematical Biology, PIMS, Canada

Abstract. We describe some genome assembly problems as a general problem of covering a hypergraph by linear and circular walks, where vertices represent sequence elements, repeated sequences are modelled by assigning a multiplicity to vertices, and edges represent co-localization information. We show that deciding if a given assembly hypergraph admits an assembly is fixed-parameter tractable, and we provide two exact polynomial time algorithms for clearing ambiguities caused by repeats.

1 Introduction

Some genome assembly problems can be seen as hypergraph covering problems, where vertices represent genomic sequences, and weighted edges encode the co-localization of sequence elements; a cover of the hypergraph with a set of linear walks (or circular walks, for genomes with circular chromosomes) corresponds to a genome assembly that respects the co-localization information encoded in the traversed edges. *Repeats*, genomic elements that appear in several locations in the genome being assembled, often confuse assembly algorithms and introduce ambiguity in assemblies. Repeats can be modelled in graph theoretical models of genome assembly by associating a *multiplicity* to each vertex, an upper bound on the number of occurrences of this vertex in linear/circular walks that cover the hypergraph. A vertex with multiplicity greater than 1 can then belong to several walks. The general assembly problem we consider is to extract a maximum weight subset of edges such that there exists a set of linear and/or circular walks on the resulting graph that contains every edge as a subwalk and respects the multiplicity of the vertices. Recent investigations describe both hardness and tractability results for related decision and optimization problems [1, 6, 2, 5].

We formalize these problems in terms of *covering of assembly hypergraphs* by linear and circular walks, and edge-deletion problems. We show that deciding if a given assembly hypergraph admits a covering by linear (circular) walks that respects the multiplicity of all vertices is FPT. We also describe polynomial time algorithms for decision and edge-deletion problems for certain instances of the problems which consider information allowing us to clear ambiguities due to repeats. Full proofs and details for each result are available in [3].

2 Preliminaries

Definition 1. *An* assembly hypergraph *is a quadruple* (H, w, c, o) *where* $H = (V, E)$ *is a hypergraph and* w*,* c*,* o *are mappings:* $w : E \to \mathbb{R}$*,* $c : V \to \mathbb{N}$*,* $o : E \to V^*$ *where* $o(\{v_1, \dots, v_k\})$ *is either a sequence on the alphabet* $\{v_1, \dots, v_k\}$ *where each element appears at least once, or* λ *(the empty sequence).*

We use the notation $|V| = n$, $|E| = m$, $s = \sum_{e \in E} |e|$, $\Delta = \max_{e \in E} |e|$, $\delta = \max_{v \in V} |\{e \in E \mid v \in e\}|$, $\gamma = \max_{v \in V} c(v)$. We call $c(v)$ the *multiplicity* of v. A vertex v s.t. $c(v) > 1$ is called a *repeat*; $V_R \subseteq V$ is the set of repeats and $\rho = |V_R|$. Edges s.t. $|e| = 2$ are called *adjacencies*; w.l.o.g., we assume that $o(e) = \lambda$ if e is an adjacency. Edges s.t. $|e| > 2$ (resp. $|e = 3|$) are called *intervals* (resp. *triples*). We denote the set of adjacencies (resp. weights of adjacencies) by $E_A \subseteq E$ (resp. w_A), and the set of intervals (resp. weights of intervals) by $E_I \subseteq E$ (resp. w_I). An interval is *ordered* if $o(e) \neq \lambda$; an assembly hypergraph with no ordered interval is *unordered*. Unless explicitly specified, our assembly hypergraphs are unordered. An assembly hypergraph with no intervals, i.e. $\Delta = 2$, is an *adjacency graph*. Given an assembly hypergraph $\mathcal{H} = (H = (V, E), w, c, o)$, we denote its *induced adjacency graph* by $\mathcal{H}_A = (H_A = (V, E_A), w_A, c, o_A)$.

Definition 2. *Let* $(H = (V, E), w, c, o)$ *be an assembly hypergraph and* P *(resp.* C*) be a linear (resp. circular) sequence on the alphabet* V*. An unordered interval* e *is* compatible *with* P *(resp.* C*) if there is a contiguous subsequence of* P *(resp.* C*) whose content is* e*. An ordered interval* e *is compatible with* P *(resp.* C*) if there exists a contiguous subsequence of* P *(resp.* C*) equal to* $o(e)$ *or its mirror.*

Definition 3. *An assembly hypergraph* $(H = (V, E), w, c, o)$ *admits a* linear assembly *(resp.* mixed assembly*) if there exists a set* $\mathcal{A}$ *of linear (resp. circular) sequences on* V *such that every edge* $e \in E$ *is compatible with* $\mathcal{A}$*, and every vertex* v *appears at most* $c(v)$ *times in* $\mathcal{A}$*. The weight of an assembly is* $\sum_{e \in E} w(e)$*.*

In the following, we consider two kinds of algorithmic problems, a decision problem and an edge-deletion problem.

- The *Assembly Decision Problem*: Given an assembly hypergraph $\mathcal{H} = (H, w, c, o)$ and a genome model (linear or mixed), does there exist an assembly of $\mathcal{H}$ in this model?
- The *Assembly Maximum Edge Compatibility Problem*: Given an assembly hypergraph $\mathcal{H} = (H = (V, E), w, c, o)$ and a genome model, compute a maximum weight subset E' of E such that the assembly hypergraph $\mathcal{H}' = (H' = (V, E'), \{w(e) \mid e \in E'\}, c, \{o(e) \mid e \in E'\})$ admits an assembly in this model.

Definition 4. *Let* $(H = (V, E), w, c, o)$ *be an assembly hypergraph. A* maximal repeat cluster *is a connected component of the hypergraph* $(V_R, \{e \cap V_R \mid e \in E\})$*.*

We now summarize some known results. Theorem 1 below follows from the equivalence between the Assembly Decision Problem with no repeats and the classical Consecutive Ones Property [4].

Theorem 1. *The Assembly Decision Problem can be solved in $O(n+m+s)$ time and space when $\gamma = 1$, in the linear and mixed genome models.*

Theorem 2. [6] *(1) The Assembly Decision Problem can be solved in time and space $O(n+m+s)$ for adjacency graphs in the linear and mixed genome models. (2) The Assembly Decision Problem is NP-hard in the linear and the mixed genome models if $\Delta \geq 3$ and $\gamma \geq 2$.*

Theorem 3. [2] *The Assembly Decision Problem can be solved in polynomial time and space in the linear genome model for unordered assembly hypergraphs where, for every edge e containing a repeat, either e is an adjacency, or e is an interval that contains a single repeat r and there exists an edge $e' = e \setminus \{r\}$.*

Theorem 4. [5] *(1) The Assembly Maximum Edge Compatibility Problem can be solved in polynomial time and space in the mixed genome model for adjacency graphs. (2) The Assembly Maximum Edge Compatibility Problem is NP-hard in the mixed genome model if $\Delta \geq 3$, even if $\gamma = 1$.*

3 New results

We now describe three positive algorithmic results, together with the corresponding algorithms and proof outlines. We first show that the Assembly Decision Problem is FPT with respect to parameters Δ, δ, γ and ρ.

Theorem 5. *The Assembly Decision Problem can be solved in space $O(n+m+s+\rho\gamma)$ and time $O\left((\delta(\Delta+\rho\gamma))^{2\rho\gamma}(n+m+s+\rho\gamma)\right)$ in the linear and mixed genome models for unordered assembly hypergraphs.*

As we consider a decision problem on unordered graphs, we omit w and o from the notation. The idea is to consider a set of derived assembly hypergraphs $\mathcal{H}_f = (H_f = (V_f, E_f), c_f)$ s.t. $c_f(v) = 1$ for all $v \in V_f$ by making $c(r)$ copies of each $r \in V_R$ and considering each possible set f of choices of 2 neighbours for each of these copies. A given $\mathcal{H}_f$ can then be checked for the existence of an assembly using Theorem 1, and $\mathcal{H}$ admits an assembly if and only if there exists an $\mathcal{H}_f$ which admits an assembly for some f. Finally, if Δ, δ, γ and ρ are bounded, there is a fixed number of such sets f, which results in an FPT algorithm.

Algorithm

1. For each $r \in V_R$, make $c(r)$ distinct copies of r. Call this set $R'(r)$, and $R' = \bigcup_r R'(r)$.
2. For each $v \in R'(r)$, choose 2 neighbours from $N'(r)$, the union of the set of non-repeat neighbours of r and of $\bigcup_p R'(p)$, the union being taken over all repeat neighbours p of r. Call this choice of neighbours for r f_r, and $f = \bigcup_{r \in V_R} f_r$.

3. Construct a new assembly hypergraph $\mathcal{H}_f = (H_f = (V_f, E_f), c_f)$ with $V_f = (V \setminus V_R) \cup R'$, $c_f(v) = 1$ for all $v \in V_f$, and E_f defined as follows: (1) for each $v_r \in R'(r)$, $r \in V_R$, $f(v_r) = \{u, v\}$ for some $u, v \in N'(r)$, add $\{v_r, u\}$ and $\{v_r, v\}$ to E_f, and (2) for each $e \in E$, add an edge $e' \in E_f$ containing $\{v \mid v \in e \setminus V_R\}$.
4. For each $v \in V_f \setminus R'$ adjacent to a vertex $r_1 \in R'$, let $v.r_1. \; \ldots \; .r_k.u$ be the unique path in H_f s.t. $\{r_1, \ldots, r_k\} \subseteq R'$ and $u \in V_f \setminus R'$. Add all of $\{r_1, \ldots, r_k\}$ to e' for each $e' \in E_f$ such that $v \in e'$.
5. Use Theorem 1 on $\mathcal{H}_f$. Output Yes and exit if $\mathcal{H}_f$ admits an assembly in the chosen genome model.
6. Iterate over all possible sets of neighbour choices f in Step 2.
7. Output No (no $\mathcal{H}_f$ admits an assembly in the chosen genome model).

We now describe an algorithm to find a maximum weight subset $S \subseteq E_I$ for an assembly hypergraph $\mathcal{H} = (H = (V, E \setminus S), w, c, o)$ to admit a mixed assembly, given that $\mathcal{H}_A$ admits a mixed assembly. We extend the notion of compatibility for an assembly hypergraph $\mathcal{H}$ as follows.

Definition 5. *An unordered interval $e \in E_I$ is said to be* compatible *with $\mathcal{H}_A$ if there exists a walk in $H_A = (V, E_A)$ whose vertex set is exactly e.*

Then, we can state the following theorem.

Theorem 6. *Let $\mathcal{H} = (H = (V, E), w, c, o)$ be an unordered weighted assembly hypergraph such that $\mathcal{H}_A$ admits a mixed genome assembly, and each interval is a triple compatible with $\mathcal{H}_A$, containing at most one repeat. Then, we can find a maximum weight subset $S \subseteq E_I$, such that $\mathcal{H}' = (H' = (V, E' = E_A \cup S), \{w(e) \mid e \in E'\}, c, \{o(e) \mid e \in E'\})$ admits a mixed assembly, in linear space and $O((n+m)^{3/2})$ time.*

The proof relies on the following ideas: (1) repeat-free triples, as well as triples whose non-repeat vertices form an adjacency, must always be included in a maximum weight compatible set of triples, and (2) the remaining triples to include can be decided using the adjacency compatibility algorithm of [5].

Algorithm

1. Initialize empty sets S, D, and $E' = E_A$.
2. Add to S every $e \in E_I$ having no repeats, and every $e = \{v_0, v_1, r\} \in E_I$ having one repeat r s.t. $\{v_0, v_1\} \in E_A$. Let $E'_I = E_I \setminus S$.
3. For every $e = \{v_0, v_1, r\} \in E'_I$ containing exactly one repeat r:
   (i) Add an adjacency $a_e = \{v_0, v_1\}$ to D. Set $w_D\,(a_e) = w\,(e)$.
   (ii) Remove $\{v_0, r\}$ and $\{v_1, r\}$ from E', if present.
4. For every adjacency $e \in E' \setminus D$, set $w'\,(e) = 1 + \sum_{a_e \in D} w_D\,(a_e)$.
5. Apply the linearization algorithm [5] (Theorem 4(1)) on $(H_D = (V, E' \cup D), w' \cup w_D, c, o_A)$.
6. For each $a_e \in D$ retained in step 5, add the triple e to S.

We finally turn to instances with larger, but ordered, intervals.

Definition 6. *Let $(H = (V, E), w, c, o)$ be an assembly hypergraph. An interval $e \in E_I$ is an* ordered repeat spanning interval *for a maximal repeat cluster R if $e = \{u, v, r_1, \ldots, r_k\}$ with $c(u) = c(v) = 1$, $\{r_1, \ldots, r_k\} \subseteq R$, and $o(e) = u.s.v$, where s is a sequence on the set $\{r_1, \ldots, r_k\}$, containing every element at least once. The subset of ordered repeat spanning intervals in E_I is denoted by E_{rs}*

Theorem 7. *Let $\mathcal{H} = (H = (V, E), w, c, o)$ be an assembly hypergraph such that every repeat $r \in V_R$ is either contained in an adjacency, or is contained in an interval $e \in E_I$ s.t. either e is an ordered repeat spanning interval, or r is the only repeat in e and there exists an edge $e' = e \setminus \{r\}$. The Assembly Decision Problem in the linear genome model can be solved for $\mathcal{H}$ in polynomial time and space.*

The basic idea of the proof is to realize the sequence $o(e)$ for every repeat spanning interval $e \in E_{rs}$ by creating unique copies of the repeats in e and decreasing the multiplicity accordingly. This leads to an assembly graph that can then be checked using Theorem 3. As we consider a decision problem, we omit w from now.

Algorithm

```
1. Let $V' = V$, $E' = E \backslash E_{rs}$, $c' = c$, $o' = \{o(e) \mid e \in E'\}$, $D = \emptyset$.
2. For every repeat spanning interval $e \in E_{rs}$:
   (a) Let $o = o(e) = u.r_1.\ldots.r_k.v$, possibly $r_i = r_j$ for $i \neq j$ (the $r_i$ are repeats)
   (b) For $i$ from 1 to $k$ add a vertex $t_i$ to $V'$, with multiplicity $c'(t_i) = 1$, and decrease $c'(r_i)$ by 1.
   (c) For $i$ from 1 to $k - 1$ add an adjacency $\{t_i, t_{i+1}\}$ to $E'$.
   (d) Add edges $\{u, t_1\}$ and $\{v, t_k\}$ to $E'$.
   (e) If the adjacencies $\{u, r_1\}$ and $\{r_k, v\}$ are present, add them to $D$.
3. Return Yes if the assembly hypergraph $\mathcal{H}' = (H' = (V', E' \backslash D), c', o')$ admits a linear genome assembly (checked by using Theorem 3) and if $c'(r) \geq 0$ for all $r \in V_R$. Else, return No.
```

References

1. S. Batzoglou and S. Istrail. Physical mapping with repeated probes: The hypergraph superstring problem. In *CPM*, volume 1645 of *LNCS*, pages 66–77, 1999.
2. C. Chauve, J. Manuch, M. Patterson, and R. Wittler. Tractability results for the consecutive-ones property with multiplicity. In *CPM*, volume 6661 of *LNCS*, pages 90–103, 2011.
3. C. Chauve, M. Patterson, and A. Rajaraman. Hypergraph covering problems motivated by genome assembly questions. arXiv:1306.4353 [cs.DS], 2013.
4. M. Dom. Algorithimic aspects of the consecutive-ones property. *Bull. EATCS*, 98:27–59, 2009.
5. J. Manuch, M. Patterson, R. Wittler, C. Chauve, and E. Tannier. Linearization of ancestral multichromosomal genomes. *BMC Bioinformatics*, 13(Suppl. 19):S11, 2012.
6. R. Wittler, J. Manuch, M. Patterson, and J. Stoye. Consistency of sequence-based gene clusters. *J. Comput. Biol.*, 18:1023–1039, 2011.